Shanghai Jiao Tong University

CS280 Elements of Data Processing

Instructor Information	TBD		
Term	March 22, 2021 - June 11, 2021	**Credits**	4 units
Class Hours	Once per week, 200 mins per teaching day		
Discussion Sessions	2 hours each week, conducted by teaching assistant(s)		
Total Contact Hours	66 contact hours (1 contact hour = 45 mins, 3000 mins in total)		
Required Texts (with ISBN)	Recommended texts: J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012. ISBN: 978-0-12-381479-1. Bing Liu, Web Data Mining, Springer, 2011. ISBN: 978-3-642-26891-5.		
Prerequisite	Students are expected to have completed one of computer programming courses such as Python, C++, Java, C#, etc. or have good knowledge of one of such programming languages.		

Course Overview

This course covers both theoretical foundations and practical techniques and tools for data processing. Topics include data representation, cleaning, transformation and analysis, visualization, privacy, clustering and classification methods, information retrieval, data and web mining, model evaluation.

Learning Outcomes

The students will be able to:

1. Have a fundamental understanding on data, data representation and storage, processing, visualization, and management.
2. Identify and use current data processing techniques, skills, and tools to perform effective data processing and analysis.
3. Have a basic knowledge of information retrieval, data mining, recommender systems, and model evaluation.

Program Outcomes

This course addresses the following program outcomes:

- An ability to apply knowledge of computing and mathematics appropriate to the discipline
- An ability to analyze a problem, and identify and define the computing requirements appropriate to its solution
- An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs
- An ability to use current techniques, skills, and tools necessary for computing practice
- The capability for critical and independent thinking and skills for lifelong learning
- Respect for academic integrity and the ethics of scholarship

Grading Policy

Participation	5%
Quizzes	5%
Presentation	10%
Staged Project	30%
Midterm	20%
Final Exam	30%

Grading Scale is as follows

Number grade	**Letter grade**	**GPA**
90-100	A	4
85-89	A-	3.7
80-84	B+	3.3
75-79	B	3
70-74	B-	2.7
67-69	C+	2.3
65-66	C	2
62-64	C-	1.7
60-61	D	1
≤59	F (Failure)	0

Class Schedule

Date	Lecture	Readings
Week 1	Why Processing Data, Data Representation, Type of Attributes, Basic Statistical Description of Data Data Integration and Cleaning: Missing Values and Outlier Detection and Removal	HKP: 3.1, 2.1-2.2 HKP: 3.2, 12.1-12.2
Week 2	Transformation by Normalization, Discretization by Binning Data Dimension Reduction	HKP: 3.5.1-3.5.3 HKP: 3.4
Week 3	Text Preprocessing and Information Retrieval Query languages and processing	L: 6.1-6.3, 6.5-6.6
Week 4	Entropy and Information Gain Association Rules	HKP: 8.2.2 L: 2.1-2.2
Week 5	Data Visualization, Clustering and Clustering Visualization Project Stage I Presentation	HKP: 2.3, 10.1-10.2 L: 4.2
Week 6	Midterm	
Week 7	Classification Methods: Decision Trees, K-Nearest Neighbor Classification Methods: Naïve Bayes, Combining Classifiers	HKP: 8.2, 9.5.1 L: 3.9 HKP: 8.3
Week 8	Experimental Design and Evaluations Link Analysis & Social Network Analysis	HKP: 8.5.1-8.5.5 L: 6.4 L: 7.1
Week 9	PageRank Assessing Correlations and Recommender Systems	L: 7.3 HKP: 2.4.7 L:12.4
Week 10	Data Preprocessing and Web Usage Mining Data Linkage, Privacy and Bloom Filters, Social and Ethical Implications of Big Data Analytics, Cloud Computing Project	L: 12.1-12.3 HKP: 13.4
Week 11	Project Stage II Presentation	
Week 12	Final Exam	